Protein Science (2000), *9*:1576–1588. Cambridge University Press. Printed in the USA.
Copyright © 2000 The Protein Society

Use of residue pairs in protein sequence–sequence and sequence–structure alignments

JONGSUN JUNG AND BYUNGKOOK LEE

Laboratory of Molecular Biology, Division of Basic Sciences, National Cancer Institute,
National Institutes of Health, Bldg. 37, Rm. 4B15, 37 Convent Drive MSC 4255, Bethesda, Maryland 20892

(RECEIVED May 20, 1999; FINAL REVISION April 28, 2000; ACCEPTED June 5, 2000)

Abstract

Two new sets of scoring matrices are introduced: H_2 for the protein sequence comparison and T_2 for the protein sequence–structure correlation. Each element of H_2 or T_2 measures the frequency with which a pair of amino acid types in one protein, k-residues apart in the sequence, is aligned with another pair of residues, of given amino acid types (for H_2) or in given structural states (for T_2), in other structurally homologous proteins. There are four types, corresponding to the *k*-values of 1 to 4, for both H_2 and T_2. These matrices were set up using a large number of structurally homologous protein pairs, with little sequence homology between the pair, that were recently generated using the structure comparison program SHEBA.

The two scoring matrices were incorporated into the main body of the sequence alignment program SSEARCH in the FASTA package and tested in a fold recognition setting in which a set of 107 test sequences were aligned to each of a panel of 3,539 domains that represent all known protein structures. Six procedures were tested; the straight Smith-Waterman (SW) and FASTA procedures, which used the Blosum62 single residue type substitution matrix; BLAST and PSI-BLAST procedures, which also used the Blosum62 matrix; PASH, which used Blosum62 and H_2 matrices; and PASSC, which used Blosum62, H_2, and T_2 matrices. All procedures gave similar results when the probe and target sequences had greater than 30% sequence identity. However, when the sequence identity was below 30%, a similar structure could be found for more sequences using PASSC than using any other procedure. PASH and PSI-BLAST gave the next best results.

Keywords: fold recognition; pair-to-pair; score matrix; sequence alignment; threading

Protein sequences are usually aligned using a scoring scheme that measures the frequency of substitution of single amino acid residues (Dayhoff et al., 1978; Gonnet et al., 1992; Henikoff & Henikoff, 1992; Jones et al., 1992; Overington et al., 1992). Here, we examine the effect of using the frequency of substitution of pairs of residues at a time that are up to four residues apart in sequence. The motivation for this exploration arises from the fact that, in common structural motifs such as helices and beta strands, close side-chain contacts are made between residues that are two, three, or four residues apart.

The idea of using pairs of residues from each protein compared is not new. Nakayama et al. (1988) and van Heel (1991) used compositions of neighboring amino acid pairs to compare database protein sequences. A large number of studies have been made on correlated mutations in which a pair of residues are mutated simultaneously (Göbel et al., 1994; Neher, 1994; Shindyalov et al., 1994; Taylor & Hatrick, 1994; Chelvanayagam et al., 1997; Olmea & Valencia, 1997). These latter studies are mainly concerned with residue pairs that are in contact in three-dimensional (3D) structure, but not necessarily close in sequence. In contrast, we will be interested in this paper in using pairs of residues that are close together in sequence so that their substitution frequencies may be used in conventional dynamic programming algorithms like those for single residues. Gonnet et al. (1994) considered a score matrix for substituting a consecutive pair of residues at a time. They reported some interesting properties of the resulting 400 × 400 matrix, but did not consider it to be useful in sequence comparisons because the database available at the time was too small to adequately fill the large matrix. However, they expressed the hope that, as the size of the database expands, matrices of this type could be used with amino acid pairs that are 1, 2, 3, or 4 residues apart in the sequence. We report here the results of using just this kind of matrices, although somewhat differently normalized.

To make the pair-by-pair comparisons, we prepared a set of four 400 × 400 (symmetric) matrices, H_2^k, where *k* was the sequence separation between a pair of residues and ranged from 1 to 4. Each element of these matrices represents the frequency with which a pair of residue types in one protein, separated in the sequence by

Reprint requests to: Byungkook (BK) Lee, National Institutes of Health, Rm. 4B15, Bldg. 37, 37 Concent Drive MSC 4255, Bethesda, Maryland 20892; e-mail: bkl@helix.nih.gov.

k residues, is replaced by another pair of residue types at the corresponding position in other structurally homologous proteins in a database. The database used was a set of over 10,000 structurally aligned pairs of protein domains prepared recently by using a new protein structure alignment program SHEBA (Jung & Lee, 2000).

Another set of four matrices, T_2^k, were generated from the same database. These are similar to the H_2^k matrices, except that each element represents the frequency with which a pair of residue types in one protein is aligned to a residue pair in another protein that has certain defined structural features. The structural feature of a residue was described by means of a structural profile (Eisenberg et al., 1997) that includes the secondary structural type and the polarity of the structural environment of the residue. We used 16 structural classes so that these matrices were 400×256 in dimension.

These matrices were tested in a fold recognition setting (Fischer et al., 1996; Miller et al., 1996; Di Francesco et al., 1997; Flockner et al., 1997; Karplus et al., 1997; Levitt, 1997; Marchler-Bauer & Bryant, 1997; Marchler-Bauer et al., 1997; Rice et al., 1997; Jones, 1999), in which 107 protein sequences (probe sequences) selected from the SCOP database (Murzin et al., 1995) were aligned to a large panel of proteins of known structure. The structure panel used contained 3,539 domains that represent all structures in the Protein Data Bank (PDB) database. It included structures that are similar as well as unrelated to the probe protein structures, with both high and virtually no sequence homologies. The tests were carried out using two computer programs, PASH (pair-to-pair alignment of sequence homology) and PASSC (pair-to-pair alignment of sequence–structure correlation). These are FASTA programs (Pearson & Lipman, 1988; Pearson, 1998) with a modification in the SSEARCH part of the program that runs the Smith–Waterman algorithm (Smith & Waterman, 1981). The modification was made only to make the dynamic programming algorithm to use pairs of residues rather than single residues. PASH uses both the single residue substitution Blosum62 (Henikoff & Henikoff, 1992) and H_2^k matrices. PASSC uses the T_2^k matrices in addition to the two sets of matrices used in PASH. A number of authors recently used a combination both the sequence homology and the structural profile information in fold recognition problems (Rice & Eisenberg, 1997; Jaroszewski et al., 1998; Russell et al., 1998). We find that PASH and PASSC perform similarly to FASTA when the probe and target sequences have more than 30% identity, but that they find significantly more sequences with identities below 30% that are structurally homologous.

Results

Characteristics of the score matrices H_2^k

The H_2^k matrix elements are labeled by a quartet of residue types, $RR_kR'R'_k$, where R and R_k are the residue types of a pair of residues that are k-positions apart in the sequence of one protein and R' and R'_k are the residue types of the matching residues in another protein that is aligned to the first. The matrix elements were evaluated by counting the frequency with which each quartet of residue types occurs in an aligned protein pair (APP) database. A matched pair of residue pairs $RR_k{:}R'R'_k$ contributes to both the $RR_kR'R'_k$ and the $R'R'_kRR_k$ elements of H_2^k. This makes the matrices symmetric, which means that there are only 80,200 unique elements in each matrix.

The APP database was prepared from the structural alignments of all known protein domain representatives (see Materials and methods) using the structure alignment program SHEBA (Jung & Lee, 2000). The total number of aligned quartets in the database was 892,724 so that the average number of observations per quartet was 11. However, a large number of quartets were not observed. The number of matrix elements with low counts are given in Table 1 for $k = 1$. The data for other k values are nearly the same. A similar data from Gonnet et al. (1994) are given for reference. The database of Gonnet et al. is nearly twice as large (1,743,134 aligned quartets) as our APP database. In addition, because of the way we recognize structurally homologous protein pairs (see Materials and methods), over 85% of the protein pairs in our APP database (9,306 out of 10,712) occur in duplicates (A to B and B to A). Duplicated pairs tend to be more structurally similar than those that occur only once. For the purpose of comparison with the matrix of Gonnet et al., two counts in our matrix elements are approximately equivalent to one count in those of Gonnet et al.

Table 1 shows that the Gonnet database produces more matrix elements with no observation and less number of elements with high counts, both in absolute counts and in percentages, than the APP database. This is probably because the aligned proteins in the Gonnet database are made of highly homologous proteins culled from a sequence database, whereas those in the APP database consist of structurally similar, but not necessarily sequentially homologous, proteins. That the latter database consists of less homologous proteins is also evident from the increased number of mutations observed among the aligned residue quartets (Table 2).

To see if the quartet frequencies carry more information than is available in single pair frequencies alone, the R^Aq/p ratios were computed, where $R^Aq/p = Pq(RR_k{:}R'R'_k)/[Pp(R{:}R') * Pp(R_k{:}R'_k)]$. The superscript A here is to distinguish this ratio from the "random" ratio, which will be described shortly. Pq and Pp are the probabilities of finding an aligned quartet or pair types, respectively, in the database. These were computed as the observed number of the quartet or pair types divided by the total number of aligned quartets, which was equal to the total number of aligned pairs (see Materials and methods). If aligned pairs, $R{:}R'$, occur independent of each other, i.e., neighboring aligned pairs are uncorrelated, then R^Aq/p will be equal to 1. Figure 1 shows the actual distributions of the R^Aq/p values for different k values. The considerable spread of the distributions indicates that the quartet frequencies indeed carry information that is not present in the single pair frequencies alone.

The frequencies of the aligned quartets were compared to "random" quartet frequencies. These latter frequencies were obtained by ignoring the structural alignment and counting all conceivable, rather than only the aligned, matches of RR_k type in one protein and $R'R'_k$ type in the other in each aligned pair of proteins in the APP database (see Materials and methods). The frequencies of "random" single pairs were also counted in the similar manner and, from these, R^Rq/p values were calculated for the "random" alignments. The distributions of R^Rq/p values are also shown in Figure 1. The spread observed in the distribution of "random" quartets must result from the high correlation that exists between residue types of sequence neighbors (van Heel, 1991). Some of the spread in the distribution of aligned quartets must be also due to this in-sequence pair correlations. However, the larger spread in the distribution of aligned quartets indicates that there are additional correlations between neighboring aligned pairs, over and above those between neighboring single residues.

Table 1. *Number of H_2^1 matrix elements with low counts*

From APP database			From Gonnet et al. (1994)		
Number of observations	Number of matrix elements	%	Number of observations	Number of matrix elements	%
0	15,811	19.7	0	36,674	45.7
1–2	13,021	16.2	1	13,588	16.9
3–4	9,687	12.1	2	6,870	8.6
5–6	7,609	9.5	3	4,110	5.1
7–8	5,716	7.1	4	2,809	3.5
9–10	4,524	5.6	5	1,932	2.4
11–12	3,617	4.5	6	1,506	1.9
13–14	2,981	3.7	7	1,104	1.4
15–16	2,445	3.0	8	925	1.2
17–18	2,042	2.5	9	792	1.0
19–20	1,642	2.0	10	686	0.9
>20	11,105	13.8	>10	9,204	11.5

The score matrix elements were calculated as logarithms of odds ratios, P^A/P^R, where P^A and P^R are the probabilities of finding an aligned and "random" residue quartet types, respectively, in the APP database. This is in contrast to Gonnet et al. (1994), who used the logarithm of R^Aq/p. Use of the single aligned pair frequencies as the reference is probably good for recognition. Use of "random" (unaligned) frequencies is probably better for accurate alignment. The odds ratio distributions are shown in Figure 2. For all k-values, about 68% of the matrix elements had the odds ratio below 1.0. Since we use the Smith–Waterman algorithm for alignment, wherein negative scores were set to zero, these elements did not participate in the alignment process. However, more than 65% of all aligned quartets in the database were of the type represented by one of the remaining 32% of the matrix elements for all k-values. Matrix elements with large number of observations, high R^Aq/p ratio, and also high scores (odds ratio) are listed in Table 3. All are diagonal elements of the matrices. Gly and, to a lesser extent, Asp and Pro appear to be the most important residues for correct structural alignment, since they occur most frequently in this list.

Characteristics of the score matrices T_2^k

The T_2^k matrix elements are labeled by a pair of residue types and a pair of residue environment types, $RR_kE'E'_k$, where R and R_k are the residue types of a pair of residues that are k-positions apart in the sequence of one protein and E' and E'_k are the residue environment types of the matching residues in another protein that is aligned to the first. A residue environment type is specified by two features; the secondary structural type in which the residue is found and the degree of exposure to the polar environment, as measured by the fraction of the accessible surface area of the side chain of the residue that is either exposed to the solvent or buried by a polar atom of the protein. There are four secondary structural types, designated as *h*, *s*, *t*, and *c* for helix, sheet, turn, and coil, respectively. The degree of exposure to the polar environment is also classified into four categories, designated as 0, 1, 2, and 3 for 0–25, 25–50, 50–75, and 75–100% exposure, respectively. Thus, there are 4 × 4 = 16 residue environment types and the total

Table 2. *Mutation rates*[a]

	APP	Gonnet et al.
None	48,426	1,071,219
One	256,974	506,251
Two	587,324	165,664
Total	892,724	1,743,134

[a] Number of quartets $RR_k{:}R'R'_k$ for which both aligned pairs are the same (none: $R = R'$ and $R_k = R'_k$), only one is preserved (one: either $R = R'$ or $R_k = R'_k$), or both pairs are different (two: $R \neq R'$ and $R_k \neq R'_k$). The second column is for our APP database; the third column is from Gonnet et al. (1994).

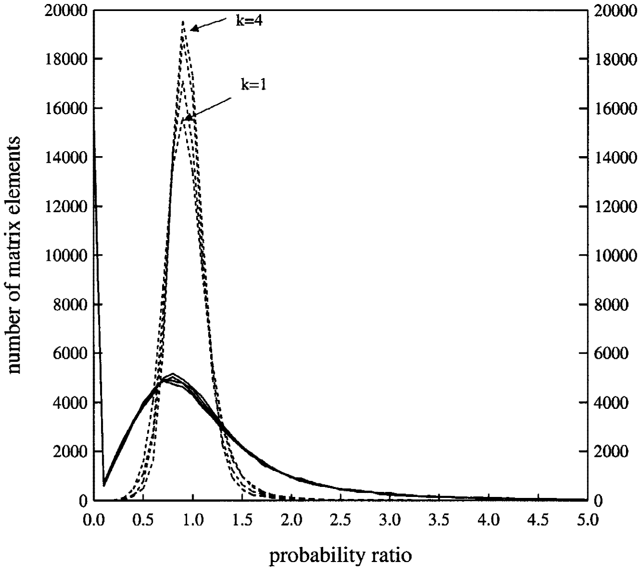

Fig. 1. Number of H_2^k matrix elements as a function of the aligned (solid line) and "random" (dotted line) quartet to single pair probability ratio, Rq/p, each for four different k values.

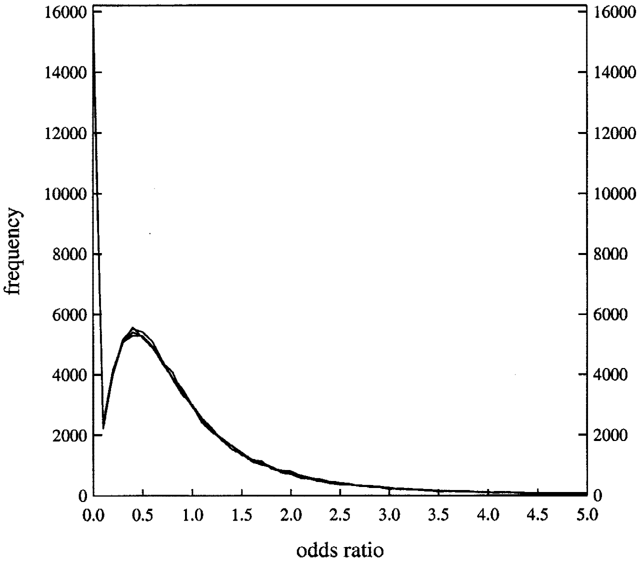

Fig. 2. Number of H_2^k matrix elements as a function of the odds ratio, P^A/P^R, for four different k values.

number of T_2^k matrix elements is $20 \times 20 \times 16 \times 16 = 102{,}400$. Each aligned quartet of residues between two proteins A and B contributes to two matrix elements: once to the element R and R_k of protein A and E and E_k of protein B and again to the element R and R_k of protein B and E and E_k of protein A. The average number of observations per matrix element was 17.

Table 3. *H_2^k matrix elements with odds ratio >20, probability ratio >2, and f^R > average frequency (2,230)*

RR_k:$R'R_k'$[a]	k[b]	Oratio[c]	Pratio[d]	f^A[e]	f^R[f]
DG:DG	2	49.1	2.9	878	3,579
GP:GP	2	48.0	2.8	736	3,072
SP:SP	3	31.5	4.6	443	2,815
GD:GD	1	31.3	2.9	890	5,688
FG:FG	3	28.5	2.0	350	2,461
PV:PV	2	27.6	2.7	432	3,130
GG:GG	2	27.4	2.1	1,470	10,737
DS:DS	1	27.3	4.9	547	4,008
LP:LP	1	26.7	2.5	510	3,827
GG:GG	3	26.6	2.1	1,442	10,844
SG:SG	2	23.2	3.1	779	6,722
DT:DT	3	21.5	2.8	303	2,825
DT:DT	1	20.4	2.8	298	2,922
DI:DI	1	20.0	2.6	312	3,124

[a] RR_k:$R'R_k'$, residue types of the four residues of a matrix element. The residue types are given in one-letter codes.

[b] k, residue separation.

[c] Oratio, odds ratio $= P_{ij}^A/P_{ij}^R$ where A = aligned, R = random, and $P_{ij} = P(RR_k{:}R'R_k')$.

[d] Pratio, probability ratio $= P_{ij}^A/(P_i^A * P_j^A)$, where A = aligned, $P_i = P(R{:}R')$, and $P_j = P(R_k{:}R_k')$.

[e] f^A, frequency observed in the aligned list. The average frequency over all types is 11.

[f] f^R, frequency observed in the "random" list. The average frequency over all types is 2,230.

The number of T_2^k matrix elements with low counts are given in Table 4. Again, because most of the protein pairs in our APP database occur in duplicates, we binned the frequency data in steps of two counts. Column 6 of Table 4 shows that more than 40% of the $k = 1$ matrix elements have not been observed. This percentage is more than twice that for the H_2 matrix, but similar to the 46% for the sequence homology matrix of Gonnet et al. (Table 1). This is understandable since the APP database is made of structurally aligned proteins; many residue pair types are never found aligned to certain environmental pair types, just as many residue pair types are never found aligned with certain other residue pair types in sequentially homologues pairs of proteins. The number of matrix elements with zero observation decreases as k increases, which indicates that the correlation between residue type pair and structural feature pair decays as the sequence separation between the pair increases.

The RR_k:$E'E_k'$ quartet probabilities were again compared to the product of the R:E' single probabilities to see if the quartets carry more information than contained in the single pairs. Figure 3 shows the distribution of the R^Aq/p and R^Rq/p ratios. The large spread of the R^Rq/p distributions reflects the strong in-sequence correlation that exists between the environments (primarily the secondary structure but probably also the polarity) of neighboring residues. The distribution becomes narrower and more centered around the unity for larger k values, as expected since the in-sequence correlation must decrease as k increases. The R^Aq/p distribution is also broad and clearly different from the "random" distribution. Thus, as in the case of the sequence homology, the quartet probabilities are different from the product of the singles and the difference is over and above that expected from the in-sequence correlation alone.

The T_2^k matrix elements, like the H_2^k matrix elements, were also calculated as logarithms of odds ratios, P^A/P^R, where P^A and P^R

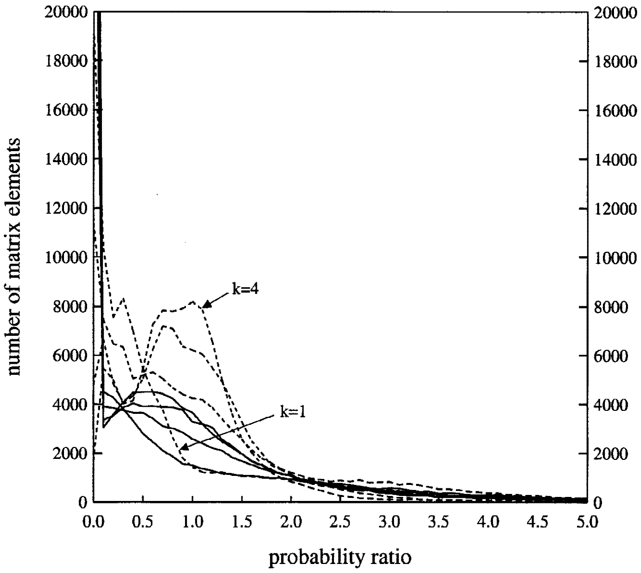

Fig. 3. Number of T_2^k matrix elements as a function of the aligned (solid line) and "random" (dotted line) quartet to single pair probability ratio, Rq/p, each for four different k values.

Table 4. *Number of T_2^k matrix elements with low counts*

Number of observations	Number of matrix elements				%			
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
0	42,601	35,138	31,227	29,562	41.6	34.3	30.5	28.9
1–2	13,594	13,962	14,032	13,764	13.3	13.6	13.7	13.4
3–4	7,706	8,516	8,855	9,018	7.5	8.3	8.6	8.8
5–6	5,148	5,882	6,187	6,440	5.0	5.7	6.0	6.3
7–8	3,743	4,369	4,761	4,889	3.7	4.3	4.6	4.8
9–10	2,862	3,440	3,767	3,857	2.8	3.4	3.7	3.8
11–12	2,291	2,838	3,089	3,325	2.2	2.8	3.0	3.2
13–14	1,983	2,381	2,582	2,699	1.9	2.3	2.5	2.6
15–16	1,634	2,037	2,193	2,292	1.6	2.0	2.1	2.2
17–18	1,420	1,772	1,989	2,045	1.4	1.7	1.9	2.0
19–20	1,240	1,503	1,709	1,825	1.2	1.5	1.7	1.8
$>$20	18,178	20,562	22,009	22,684	17.8	20.1	21.5	22.2

are the probabilities of finding an aligned and "random" pair of pairs, $RR_kE'E'_k$, respectively, in the APP database. Matrix elements with large number of observations, high R^Aq/p ratio, and also high scores (odds ratio) are listed in Table 5. Buried or partially buried Cys residue type appears most frequently in this list.

Table 5. *T_2^k matrix elements with odds ratio $>$15, probability ratio $>$2, and $f^R >$ average frequency (3,493)*

RR_k:$E'E'_k$[a]	k[b]	Oratio[c]	Pratio[d]	f^A[e]	f^R[f]
AC:c2h1	3	40.6	28.3	723	3,567
HC:h2h1	1	32.7	98.0	846	5,187
AH:c2h2	2	27.4	18.5	696	5,085
VA:s0h1	4	22.4	2.2	419	3,742
LC:s0s0	2	21.5	18.6	1032	9,623
DS:c1t2	1	21.1	30.1	657	6,237
YC:s1s0	2	20.2	20.8	756	7,507
AC:s1s0	4	19.4	8.0	371	3,839
DG:c1t1	2	17.4	26.6	310	3,562
CV:s0s0	2	17.3	14.0	802	9,296
TC:s2s0	3	16.8	9.4	522	6,211
GW:s2c1	1	16.0	19.9	333	4,167
CD:c1c1	3	15.9	30.6	526	6,636
VW:s1c1	3	15.8	6.9	329	4,166
WG:s1t3	4	15.6	5.8	400	5,148

[a] RR_k:$E'E'_k$, two residue types (RR_k) in one protein and a pair of environment types $(E'E'_k)$ in the other, matched protein. The residue types are given in one-letter amino acid codes. The environment E is defined by a combination of a secondary structural type and the degree of exposure to solvent and other polar environment. For secondary structural elements: h, helix; s, β-sheet; t, turn; c, coil. For the degree of exposure to a polar environment: 0, 0–25% exposure; 1, 25–50%; 2, 50–75%; 3, 75–100%.

[b] k, residue separation.

[c] Oratio, odds ratio $= P_{ij}^A/P_{ij}^R$ where $A =$ aligned, $R =$ random, and $P_{ij} = P(RR_k{:}E'E'_k)$.

[d] Pratio, probability ratio $= P_{ij}^A/(P_i^A * P_j^A)$, where $A =$ aligned, $P_i = P(R{:}E')$, and $P_j = P(R_k{:}E'_k)$.

[e] f^A, frequency observed in the aligned list. The average frequency over all types is 17.

[f] f^R, frequency observed in the "random" list. The average frequency over all types is 3,439.

Entropy of the score matrices

Given a score matrix, the average score per alignment S is given by $\sum P_{ij}^A s_{ij}$, where P_{ij}^A is the probability of an alignment, s_{ij} is an element of the score matrix, and the summation is over all the matrix elements. When the score matrix is defined as base 2 logarithm of the odds ratio, S is also the relative information theoretical entropy of the target alignment in bit units (Altschul, 1991; Karlin & Altschul, 1991). The entropy values for the H_2^k and T_2^k matrices are 0.70 and 0.50, respectively, for all k values. The value for the H_2^k matrices is similar to those of PAM160 (Dayhoff et al., 1978; Altschul, 1991) and Blosum62 (Henikoff & Henikoff, 1992, 1993) matrices. The fact that S is less for the T_2^k than for the H_2^k matrices indicates that sequence–environment correlation is less than sequence–sequence correlation even for the structurally aligned APP database. The entropy values for the H_1 and T_1 matrices were also calculated where H_1 and T_1 matrices are the single residue pair and single residue–environment pair alignment score matrices, respectively, calculated using the same APP database. They are 0.25 and 0.15, respectively. The entropy of the H_1 matrix is comparable to that of the PAM310 matrix and between those for the Blosum35 and Blosum40 matrices. The low value again indicates that protein pairs in our APP database are not highly sequentially homologous. The fact that the entropies for the H_2^k and T_2^k matrices are significantly higher than those H_1 and T_1 states that there is a large increase in information content when pairs of pairs are used compared to the single pairs.

Number of correct hits and false positives

Figure 4 shows the z-scores obtained when each of the 107 probe sequences was aligned to every protein in the 3,539-protein domain database using the PASSC procedure. The z-scores are those used by the FASTA program and correspond to the ordinary z-score times 10 plus 50. The maximum z-score was 2,691 for the 1gfn–2omf pair, which had 100% sequence identity. The minimum was 20 for the 1pdo–1ksiA0 pair, which were not structurally related and had 0% sequence identity after the alignment. The figure shows only those with z-scores between 50 and 300. As can be seen from the plot, the z-score distributions do not show significant dependence on the size of the probe. This was achieved because the

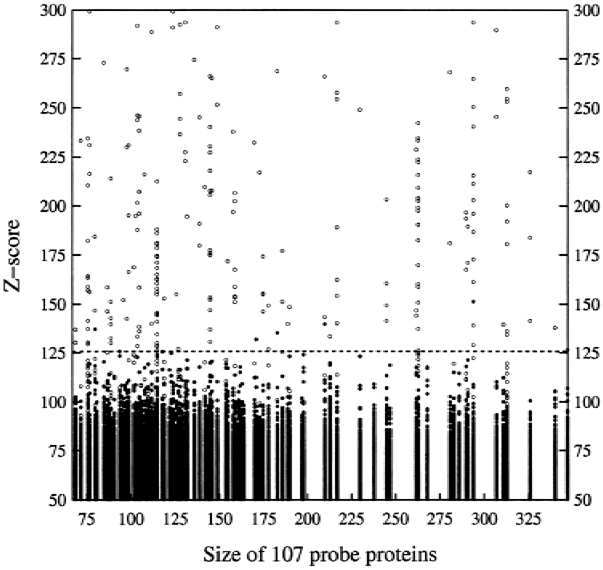

Fig. 4. Z-cores for the 107 probe sequences when each were optimally aligned to all of the 3,539 proteins in the domain database using the PASSC procedure. The z-scores plotted are those used in FASTA, which correspond to the true z-score times 10 plus 50. Z-scores below 50 or above 300 are not shown. Open and closed circles are for the structurally homologous and nonhomologous pairs, respectively, according to the structure–structure alignment by SHEBA. The horizontal dotted line indicates the z-score cutoff value used in Table 6.

z-scores were computed using a facility in FASTA that corrects for the size of the proteins (Pearson, 1998) and by the judicious choice of the weights in Equation 3 in Materials and methods. A similar lack of dependence on the size of the probe sequence was seen when the probes were aligned using the unmodified FASTA, Smith–Waterman (SW), or the PASH procedures. It is also clear from the figure that the pairs with high z-scores are structurally homologous and that structurally nonhomologous pairs begin to appear as the z-score is lowered.

To see how the structurally homologous and nonhomologous pairs ranked in terms of the z-scores, the 107 ∗ 3,539 aligned protein pairs were sorted in descending order of their z-scores. The cumulative number of structurally homologous pairs, with less than 30% sequence identity between the pair, were counted and plotted as a function of the rank in the sorted list. The results are shown in Figure 5 for the four different alignment procedures. The figure shows that, in terms of the number of structurally homologous protein pairs that occur among the top-scoring protein pairs of low sequence homology, PASH is indeed clearly better than either FASTA or SW and that PASSC makes a further improvement over PASH.

Another way to measure the performance of the different alignment procedures is to count the structurally homologous (correct hits) as well as the nonhomologous (false positives) pairs that have z-scores above a cutoff value. Obviously, if the z-cutoff value is set high, there will be no false positives, but the number of correct hits will be small. Lowering the z-cutoff value increases the number of correct hits but the number of false positives also increases. We chose the z-cutoff value as the lowest z-score that still maintains the number of false positives to less than 2% of the total number of pairs above the cutoff value. The numbers of correct hits and false positives with this choice of z-cutoff are given in Table 6. Since pairs with high sequence homology tend to have high z-scores regardless of the alignment procedure used, we counted the number of correct hits separately for those with less or higher than 30% sequence identity. There were 99 pairs with 100% sequence identity; these were not counted. Column 3 shows that all alignment procedures find essentially the same number of correct hits when the sequence identity is higher than 30%. The number of correct hits with less than 30% sequence identity is also the same between the FASTA and SW procedures (Column 4), but it increases by 50% when the PASH procedure is used and more than doubles when the PASSC procedure is used.

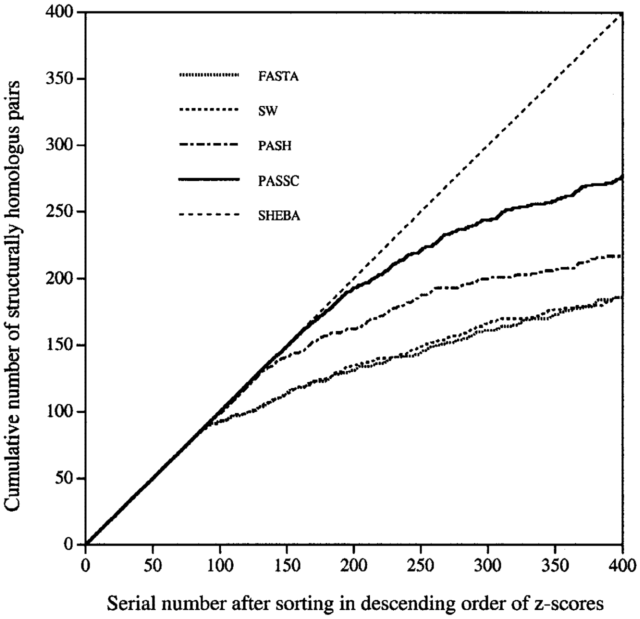

Fig. 5. Cumulative number of structurally homologous pairs as a function of the rank in the list of all pairs sorted in descending order of the z-score. The protein pairs and their z-scores are for the alignments of 107 probe sequences to the 3,539 domains using the FASTA, SW, PASH, and PASSC procedures as indicated, except that the pairs that have more than 30% sequence identity after structural alignment by SHEBA were omitted. The Figure shows only the 400 highest scoring pairs. If all structurally homologous pairs had z-scores higher than any structurally nonhomologous pair, the diagonal line is expected, since the total number of structurally homologous pairs is much more than 400 (see Table 6). The deviation of each curve from the diagonal gives the cumulative number of structurally nonhomologous pairs.

The total number of structurally homologous pairs with less than 30% sequence homology is 3,339. Thus, the actual number of correct hits obtained using any procedure is only a small fraction of the possible total, indicating that a large number of structurally homologous pairs with low sequence homology have z-scores below the cutoff value for all procedures. It can also be noted that, even among those with greater than 30% sequence identity, the number of correct hits is less than the possible total. The eight pairs with better than 30% identity, but which did not register as the correct hits by PASSC alignment, were examined. It was found that the alignments were essentially the same as the SHEBA alignments for all eight cases. However, for seven of the eight cases, the target domain was one-third to one-fourth the size of the probe. Thus, even though each of these domains was a true structural

Table 6. *Results after threading 107 chains through the 3,539 domains*[a]

		Number of correct hits			Number of probes with at least one correct hit	
	z-cutoff	%id[c] ≧ 30	%id[c] < 30	N_f[b]	%id[c] ≧ 30	%id[c] < 30
FASTA	128.0	115	84	2	44	46
SW	132.6	115	84	2	44	45
PASH	125.5	116	125	1	44	55
PASSC	125.7	114	183	5	43	62
SHEBA		122[d]	3,339[d]		46[e]	107[e]

[a] Pairs with 100% identity were omitted.
[b] Number of false positives.
[c] Percentage of identical residues among the aligned residues.
[d] Total number of structurally homologous pairs in the specified sequence homology ranges.
[e] Total number of probe sequences which have at least one homologous structure in the domain database in the specified sequence homology ranges.

homologue of a part of the probe structure, the part was too small to score high by the FASTA criteria, which obtains scores for the global alignment. The eighth protein pair was 1bdo–1fyc, which had 33% sequence identity between them and a z-score of 121. There were three other domains that had higher z-scores, all of which were correct hits for this probe. The z-score cutoff value used was obviously too high for this probe.

Alignment shifts compared to the structure–structure alignment

The sequence–sequence and sequence–structure alignments obtained here can be compared to the structure–structure alignment by SHEBA using the average alignment shift, Δr. This latter quantity is defined as $\Delta r = \sum \Delta r_{ii'}/N$, where i and i' are two residues, one from the probe and the other from the target sequences, which are aligned in the structural alignment by SHEBA, $\Delta r_{ii'}$ is the number of residues and gaps that separate these residues in the alignment by one of the procedures described here, N is the total number of aligned residues in the SHEBA alignment, and the summation is over all N aligned residue pairs. The distribution of Δr values for alignments with less than 30% sequence identity is given for each of the three different alignment procedures in Table 7. In most cases, the alignments obtained by these procedures are essentially the same as those obtained by structural alignments, but there are also some correct hits that nonetheless have very different alignment from what SHEBA obtains. The three cases with the largest alignment shifts after PASSC alignment were examined in detail. In the case of the bacterial luciferase (1lucA–1xkjB1) and carboxypeptidase–lactamase (3pte–1blsA1) pairs, SHEBA alignments include large gaps, of 173 and 98 residues long, respectively. The PASSC procedure, being a variant of the FASTA program, uses extension gap penalty and does not allow such a large gap. Instead, it aligned the last half of the domain structure to the structurally nonhomologous middle part of the probe sequence. The average alignment shifts produced were 101 and 58, respectively. The other pair was the membrane protein porins (2omf–1prn), which have a barrel structure made of 16 β-strands. In the alignment obtained by PASSC, the barrels were rotated by two strands relative to each other when compared to the SHEBA alignment, causing an average alignment shift of 59. The percent sequence identity increased from 12 to 26% after this rotation.

Table 7. *Number of correct hits*[a] *in different alignment shift (Δr) categories*

	$\Delta r \leqq 5$	$5 < \Delta r \leqq 10$	$\Delta r > 10$
SW	64	8	12
PASH	96	16	13
PASSC	151	18	14

[a] Only those with less than 30% sequence identities were counted.

Fold recognition in different ranges of sequence homology

Not surprisingly, many correct hits were found for some probe sequences while no hit was found for some others. Since one correct hit is sufficient to identify the fold of a given probe sequence, the number of different probe sequences represented in the list of correctly hit probe-target pairs is of interest. Figure 6 shows the number of probe sequences with at least one correct hit in different ranges of percent sequence identities. It also shows the maximum possible number in each sequence range, which is the total number of probe sequences with a structural homologue in the indicated sequence identity ranges regardless of the z-score.

There were 83 probe sequences that had a correct hit with 100% sequence identity. Excluding these self-matches, the number of probe sequences with at least one correct hit is small when the sequence identity is greater than 30% because the domain database is such that no two domains have sequence identity greater than 50% (Jung & Lee, 2000). In these high sequence identity ranges, all alignment procedures find essentially the same maximum possible number of probes. The number is also small when the sequence identity is below 10%, despite the fact that most probe sequences have at least one structural homologue in these ranges. This indicates the limited power of the procedures used in this study for identifying structural homologues with little sequence homology. In the middle range of sequence identities, the number of probe sequences with at least one correct hit is clearly larger when PASH or PASSC is used than when FASTA or SW is used. For example, for the 20–25% sequence identity range, both FASTA

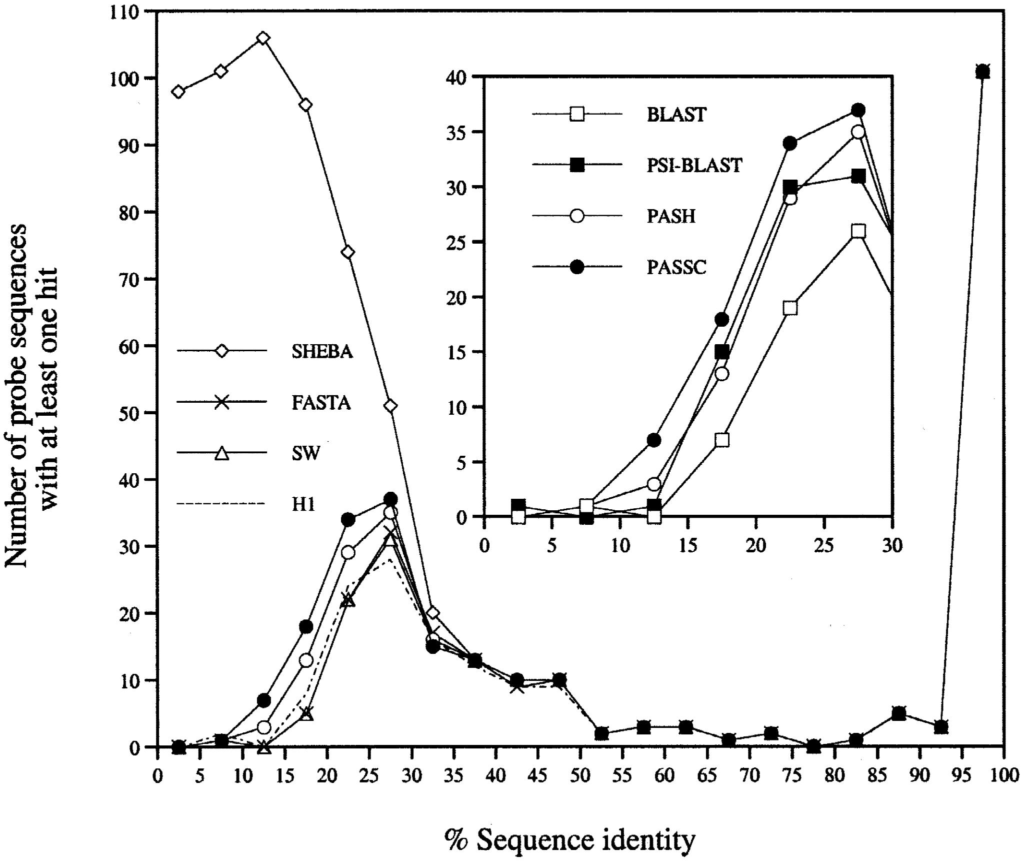

Fig. 6. Number of probe sequences that have at least one correct hit in different 5% ranges of percent sequence identity between the probe and the correct hit target sequences. The alignment procedures used are FASTA (cross), SW (triangle), FASTA using H_1 matrix (dotted line), PASH (open circle), PASSC (solid circle), BLAST (open square), and PSI-BLAST (solid square). The curves for FASTA and SW are nearly identical at all sequence identity ranges. The curve with open diamonds indicates the number of probe sequences that have at least one structurally homologous pair in the indicated sequence homology ranges, regardless of whether the z-score is above or below the cutoff value. All curves superimpose exactly when the sequence identity is more than 50%.

and SW find at least one correct hit for 22 probe sequences, compared to 29 and 34 with PASH and PASSC, respectively. The corresponding numbers for the 15–20% range are 5, 5, 13, and 18 for FASTA, SW, PASH, and PASSC, respectively. Similar data, but using only two sequence identity classes, those with higher or lower than 30% sequence identities, are also given in the last two columns of Table 6.

To see if the improvement is due to the use of the structurally aligned protein database or to the use of the pair-by-pair comparison matrices, we also ran the FASTA program using the H_1 matrix derived from APP. The result (dotted line in Fig. 6) is very similar to that using the Blosum62 matrix, although there is a hint of a better result at lower homology ranges. It appears, therefore, that the main reason for the improvement is the use of the pair-by-pair score matrices.

We also compared the PASH and PASSC procedures to the BLAST (Altschul et al., 1990) and PSI-BLAST (Altschul et al., 1997) procedures, although this comparison is not a direct test of the new score matrices since these latter procedures use a different algorithm for searching for similar sequences. As can be seen in the inset in Figure 6, PASH and PSI-BLAST perform similarly for this set of test sequences while PASSC performs noticeably better.

The recognition rates among different classes of proteins are listed in Table 8. Surprisingly, the recognition rate was the poorest for the α-class proteins, and best for the β- and "other" classes, for all alignment procedures. The structures of the α-class probe se-

Table 8. *Recognition rate for each class*

		Correct hits[b] (%)			
Class[a]	Total	FASTA	SW	PASH	PASSC
α class	21	7	7	8	9
β class	25	12	12	15	17
α/β class	29	14	14	16	17
$\alpha+\beta$ class	28	11	10	12	15
Others	4	2	2	4	4
All	107	46	45	55	62

[a] SCOP secondary structure classification (Murzin et al., 1995).

[b] Total number of probe sequences in each protein secondary structural class with at least one correct hit with less than 30% sequence identity.

quences for which PASSC found at least one correct hit are shown in Figure 7A and those for which no correct hit was found in Figure 7B. It is apparent that 10 of the 12 probe chains with no correct hit have an up-and-down helical bundle structure. It was found that the probe sequences of this type of structure shared little sequence homology with the domains that were found to be structurally homologous by structure–structure alignment using SHEBA (data not shown). The 1lpe–1cgo2 pair given in Table 9 is an example.

Table 9 shows the sequence identity and the H_{1b} (Blosum62), H_2 and T_2 terms that were used in the alignment score calculation for some sample probe-target protein pairs after they are aligned structurally by SHEBA. Two examples were drawn from each secondary structural class, a correct hit and another that did not register a z-score above the z-cutoff value. It can be seen that the pairs with low z-scores have generally low sequence identity between them and have negative H_{1b} and H_2 scores, in contrast to those that have a sufficiently high z-score to be recognized as a correct hit pair.

Discussion

There are many reasons why one performs a protein sequence alignment, but one main purpose is to identify proteins that have a similar structure to that of a given sequence. Testing a new sequence alignment procedure for this purpose is a complicated process because simply increasing the number of sequences that score significantly higher than a random match, as is usually done, is not sufficient. It must be shown in addition that the newly found sequences have a similar structure to that of the given sequence (Brenner et al., 1998) that involves comparing two structures that

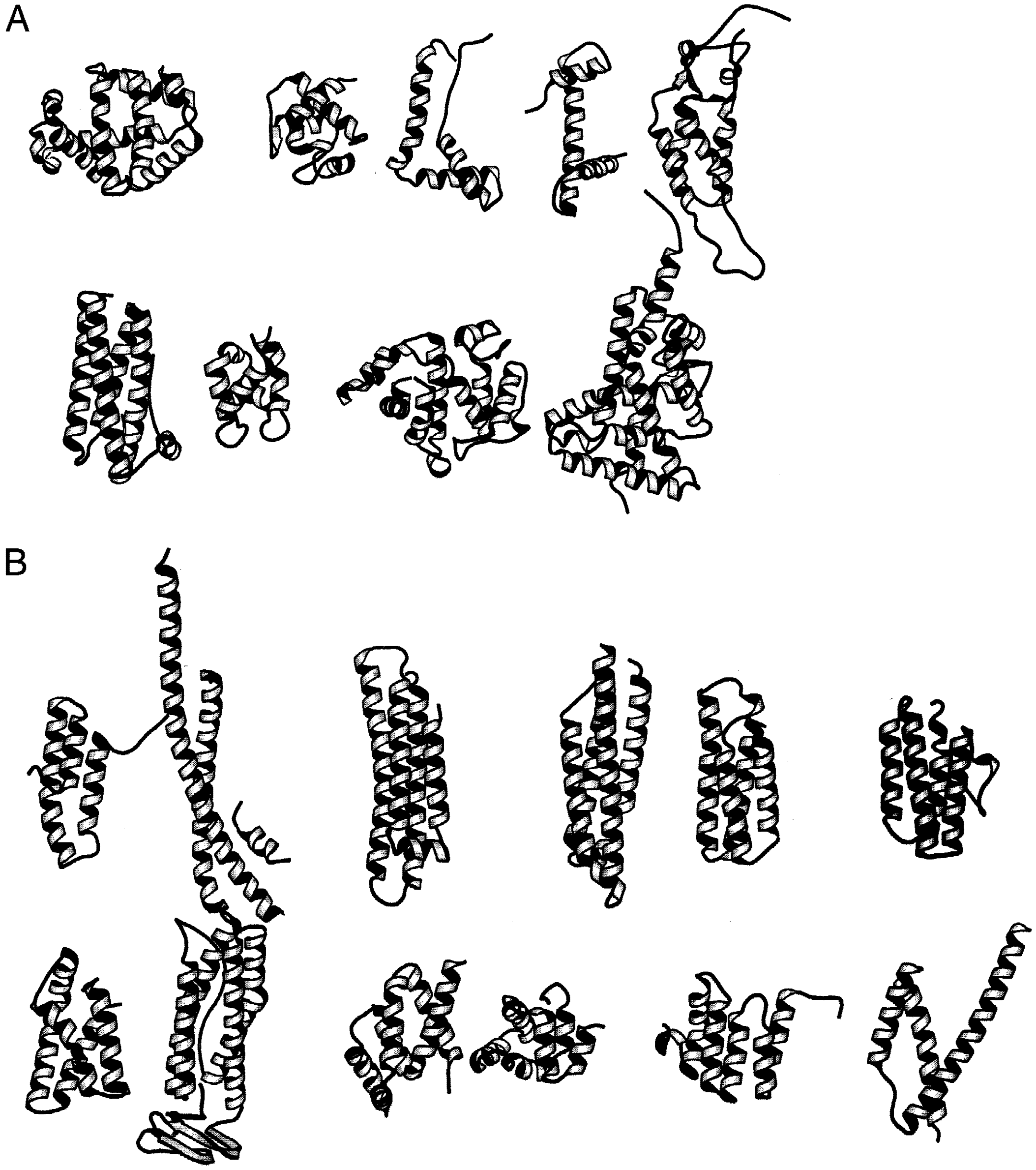

Fig. 7. MOLSCRIPT (Kraulis, 1991) drawings of the structures of the α-class probe sequences with at least (**A**) one or (**B**) no correct hit with less than 30% sequence identity.

Table 9. *Components of the alignment scores for some sample probe and target protein pairs*

PDB names	PASSC[a]	Class[b]	M[c]	ID[d]	H_{1b}[e]	H_2[e]	T_2[e]
3sdhA[f]–1eca	✓	α	90	11	35	207	198
1neu–3hfmH	✓	β	90	14	2	175	168
3chy–1ntr	✓	α/β	66	22	50	154	80
1kuh–1hfc	✓	$\alpha+\beta$	83	25	75	260	97
1lpe[g]–1cgo2	—	α	73	6	−65	−51	18
1who–1cid-2	—	β	73	12	−25	17	102
2rslA–1ah3B3	—	α/β	68	18	−6	54	22
1fwp–1mla-2	—	$\alpha+\beta$	87	5	−50	−68	19

[a] Check mark shows when PASSC recognizes the target as a correct hit.

[b] SCOP secondary structure classification (Murzin et al., 1995).

[c] Number of structurally matched residues as a percentage of the size of the smaller protein.

[d] Percent identical residues among the aligned residues after SHEBA alignment.

[e] The three score terms in Equation 3, summed over all k-values and all aligned residue pairs, calculated after the structure-structure alignment by SHEBA.

[f] This is a globin shown as the first structure in Figure 7A.

[g] This protein has the up-and-down helix bundle structure shown as the third structure in Figure 7B.

can be only partially or remotely similar. We used the structure–structure comparison program SHEBA for this purpose and used the more or less arbitrary criterion that the number of structurally matched residues be greater than 60% of the number of residues in the smaller protein. Obviously, one can use a different set of criteria with equal justification, which can substantially change the number of identified sequences. But, it seems unlikely that such a change would materially alter the relative performance of the two new alignment procedures reported herein with respect to the more traditional SW or FASTA procedure. Also, the tests we devised are just to test the new pair-by-pair matrices, rather than to assess PASH or PASSC as new threading procedures, in which case the absolute number of identified sequences would be an important consideration.

The new H_2^k and T_2^k matrices, which involve pairs of amino acids, are large matrices of dimensions 400 × 400 and 400 × 256, respectively. Setting up these matrices requires a large database of aligned pairs of proteins, which became available to us only after the fast, large scale structure–structure alignments using SHEBA (Jung & Lee, 2000). Even with 10,712 aligned protein pairs, the matrices are relatively sparse and contain many elements that are zero for lack of data. Nevertheless, these matrices contain more information than the single pair comparison matrices. This can be seen from the fact that the distribution of pair–pair probabilities is clearly different from that of the product of single-pair probabilities and from the fact that the entropies for the H_2^k and T_2^k matrices are much larger than those for the H_1 and T_1 matrices.

A problem in using these matrices is that the dynamic programming algorithm is no longer rigorous in finding the best alignment when pairs of residues are involved. We used a simple algorithm that will find the optimal alignment when there is no gap (see Materials and methods), but it will not necessarily find the best alignment when gaps are introduced. The algorithm must have worked reasonably well since the results reported in Figures 5 and 6 and in Table 6 clearly show that the procedure, along with the new matrices, does improve the ability of a sequence alignment program to recognize structurally homologous proteins. It is not clear whether a rigorous algorithm can be found. The present, simple algorithm will work better in a nongapped procedure such as BLAST than in FASTA. BLAST is probably the procedure of choice for another reason as well (see below).

The fold recognition ability varies according to the sequence homology. As can be seen from Figure 6 and Table 6, FASTA finds nearly all structurally homologous proteins when the sequence identity is better than 30%. Therefore, no improvement is made with the new procedures in this sequence homology range. However, all procedures failed to recognize some proteins with partial structural homology in this high sequence homology range. Examination of the nature of the failures indicates (see Results) that use of a local alignment procedure such as BLAST (Altschul et al., 1990, 1997), as opposed to a global alignment, would have improved the ability to find these proteins. Improvement is also not detectable when there is less than 10% sequence identity. Some of the reasons for this lack of success at low sequence homology range have been explored. At least in the case of the α-helical proteins, we found that most of the failed cases involve simple up-and-down helix bundle architecture (Fig. 7B). The structural homologies in these cases appear accidental or to have arisen by convergent evolution, since sequence homology is low and the H_{1b} and H_2 scores are actually negative (Table 9). The CASP2 (Marchler-Bauer et al., 1997) and CASP3 (Koehl & Levitt, 1999) blind protein structure prediction experiments showed that some sequences are easier to predict than others by various fold recognition procedures and that the "difficult" cases are usually those that have low or no sequence homology. It is not surprising that PASSC, being a blend of the sequence–sequence alignment program PASH and the "profile" method of fold recognition (Bowie et al., 1991), also finds it difficult to find structurally homologous proteins at the very low sequence homology ranges.

On the other hand, there is a clear improvement in the "twilight zone" of 10–30% sequence identities. There can be two reasons for this improvement. One is that our APP database consists of protein pairs that are structurally aligned but not too highly sequentially homologous. The high mutation rate observed among the aligned protein pairs in APP (Table 2) and the low entropy of only 0.25 for the H_1 matrix show this to be the case. Henikoff and Henikoff (1993) pointed out some time ago that matrices derived from structurally aligned database tend to perform better than those derived from sequentially aligned database. However, the fact that FASTA does not perform any better using the H_1 matrix than when the Blosum62 matrix is used indicates that this is not likely to be the major reason. The other, more likely, reason is that pairs of residues contain additional information not present in single residues.

Materials and methods

Domain and aligned protein pair databases

The database used to calculate the scoring matrices consisted of aligned pairs of protein domains of known structure. These were obtained as follows (Jung & Lee, 2000). There were a total of 13,983 protein chains in the March 1998 release of the PDB. These were broken into domains using the domain parsing program, PUU (Holm & Sander, 1994). After eliminating theoretical models, nonpeptides, and domains with less than 40 amino acid residues, there were 18,595 domains. The 18,595 domains were clustered into

3,539 sequentially homologous groups using a fast version of the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970) and the Gonnet scoring matrix (Gonnet et al., 1992). The smallest protein was selected from each group to represent the group. This set of 3,539 representative proteins is referred to as the domain database.

The structure–structure comparison program SHEBA was run between all pairs of these domains (Jung & Lee, 2000) and those pairs that met the following criteria were selected: (1) The number of structurally matched residues is greater than 40 in absolute number and greater than 50% of the residues in the larger protein of the pair; (2) Z-score is greater than 4.0. The z-score, z_{ab}, between a probe sequence a and a target sequence b was calculated using the number of matched residues between a and b relative to the average number of matched residues between a and all other proteins in the domain database; and (3) the number of identical residues after the structural alignment is between 10 and 40% of the matched residues. There were 10,712 pairs that met all of the above criteria. Many of these are duplicates (*a–b* and *b–a* pairs), but some pairs with less structural homology occur only once. This set of domain pairs is referred to as the aligned protein pair (APP) database.

Score function

The total alignment score between a pair of proteins was calculated as the sum of the individual contributions $a_{ii'}$ made by each pair of aligned residues, i of one sequence and i' of the other sequence, minus the opening and extension gap penalties. The contribution $a_{ii'}$ was calculated as

$$a_{ii'} = H_1(R_i; R_{i'}) \tag{1}$$

in the FASTA and Smith and Waterman (SW) alignments,

$$a_{ii'} = H_1(R_i; R_{i'}) * 0.25 + \sum_{k=1}^{n} H_2^k(R_{i-k}R_i; R_{i'-k}R_i) \tag{2}$$

in the PASH procedure, and

$$a_{ii'} = H_1(R_i; R_{i'}) * 0.20 + \sum_{k=1}^{n} H_2^k(R_{i-k}R_i; R_{i'-k}R_{i'}) + \sum_{k=1}^{n} T_2^k(R_{i-k}R_i; S_{i'-k}P_{i'-k}S_{i'}P_{i'}) * 0.6 \tag{3}$$

in the PASSC procedure. In these expressions, i and i' indicate positions of matched residues in the two sequences and R, S, and P, with various subscripts indicating the position of the residue, represent the amino acid type, secondary structural type, and polarity type of the residue, respectively. There are four values for the secondary structural type, corresponding to helix, sheet, turn, and coil. A secondary structural type was assigned to each residue using the DSSP program (Kabsch & Sander, 1983). The polarity type refers to different ranges of polarity of the environment of the residues. The latter is defined as the fraction of the accessible surface area of a residue that is exposed to solvent or buried by a polar atom (Bowie et al., 1991; Jung & Lee, 2000). Again, four values were used for the polarity type, corresponding to 0–25, 25–50, 50–75, and 75–100% polarity ranges, respectively. H_1 is the single residue homology matrix. We used the Blosum62 matrix (Henikoff & Henikoff, 1992) for H_1, since it has higher information entropy than that calculated from the APP database. Blosum62 matrix is denoted as H_{1b} to distinguish it from H_1 calculated from the APP database. H_2^k is the k-type pair-to-pair sequence homology matrix, of dimension 400 × 400. T_2^k is a k-type pair-to-pair sequence–structure correlation matrix of dimension 400 × 256. The value of n in Equations 2 and 3 was four.

To obtain the matrices H_2^k and T_2^k, two k-type residue pair–pair lists were prepared from the APP database for each k-value. A k-type residue pair is a pair of residues that are k-residues apart in a sequence. One list consisted of all the aligned pairs of k-type residue pairs (a k-type residue pair from one sequence aligned to another pair from the second sequence) in all the protein pairs in the database. The residues that were aligned to a gap were not counted in calculating the k-value for this list. The length of the list was made the same for all k-values by not counting the quartets whose first pair falls within four residues from the C-terminus of each protein. The second list was much larger and consisted of a concatenation of all $(n_p - 4) * (n_{p'} - 4)$ pairs of k-type residue pairs per each protein pair $p - p'$, where n_p and $n_{p'}$ are the numbers of residues in the proteins p and p', respectively. Occurrence of gaps was ignored in making this list. These two lists are referred to as the aligned and random (k-type) residue pair-pair lists, respectively.

The H_2^k matrix was obtained by

$$H_2^k = \ln[P^A(RR_k; R'R'_k)/P^R(RR_k; R'R'_k)], \tag{4}$$

wherein P^A and P^R were the normalized frequencies with which a residue pair of the amino acid types R and R_k is found paired to another residue pair of the amino acid types R' and R'_k in the k-type aligned and random residue pair–pair lists, respectively. Similarly, the T_2^k matrix was obtained by

$$T_2^k = \ln[P^A(RR_k; S'P'S'_kP'_k)/P^R(RR_k; S'P'S'_kP'_k)], \tag{5}$$

wherein P^A and P^R were the normalized frequencies with which a residue pair of the amino acid types R and R_k is found paired to another residue pair which have the secondary structural types S' and S'_k and polarity ranges P' and P'_k in the k-type aligned and random residue pair–pair lists, respectively.

For the purpose of calculating the entropy, the H_1 and T_1 matrices were constructed in a similar manner, but using single pair frequencies. The total number of single pairs was made the same as that of the quartets by not counting the pairs that are within four residues from the C-terminus of each protein.

Alignment procedure

Default values were used for the gap penalties and the statistical parameters as given in the FASTA program (Pearson & Lipman, 1988; Pearson, 1998). The alignment routine in the FASTA package (version 3.0t77, downloaded from the ftp site ftp://ftp.virginia.edu/pub/fasta) was replaced with the Smith–Waterman algorithm (Smith & Waterman, 1981) given in the same package and then modified to use the H_2^k and T_2^k matrices (see below). The different weight values shown in Equations 2 and 3 were chosen to make the total score similar in magnitude to those obtained when Equation 1 was used

Table 10. *107 probe proteins*

1aaj-	1afi-	1ako-	1aps-	1bdo-	1ble-	1bplB	1ctj-	1dcoA	1draA
1e2b-	1exg-	1frd-	1fwp-	1gky-	1guaB	1hcd-	1hihA	1hxn-	1lbd-
1lfaA	1lpe-	1ltsD	1lxa-	1mldA	1mrj-	1nciA	1ngr-	1occE	1pdo-
1pil-	1pyp-	1rhgA	1rtm1	1ryt-	1std-	1tig-	1tph1	1ubi-	1vid-
1vtmP	1who-	256bA	2bopA	2cpl-	2fcr-	2hmqA	2omf-	2rslA	2vik-
3cla-	3pte-	3sdhA	4rhn-	1abv-	1aihA	1aly-	1arb-	1bfmA	1bmfG
1chd-	1cus-	1dorA	1dupA	1ecmA	1fapB	1gdoA	1gmpA	1hbp-	1hcl-
1hme-	1kuh-	1lcjA	1lit-	1lre-	1lucA	1mai-	1molA	1nal1	1neu-
1nulA	1pdgA	1phr-	1pne-	1rgp-	1ris-	1rvvA	1sriA	1tfr-	1tml-
1tul-	1vhrA	1vmoA	1wba-	1zin-	2asr-	2chsA	2ctb-	2fgf-	2mcm-
2rn2-	2trxA	3chy-	3pgm-	3rubS	4icb-	7rsa-			

with the Blosum62 matrix. In the PASSC procedure, gaps were not allowed in helices and beta strands.

The dynamic programming algorithm for finding the optimum alignment had to be modified to use scores that depend on two aligned residue pairs instead of just one pair. In the forward moving Smith–Waterman algorithm (Smith & Waterman, 1981), best alignments for the subsequences 1 to $i - 1$ of one sequence and 1 to $j - 1$ of the other are known, for all possible overhang lengths, at the time the score is calculated for aligning the residues i and j. The modification consists of using this known upstream alignment to find the residue and environment types of the aligned residue pair, k-positions upstream, for each overhang. Gaps were considered like a residue in counting the k-value. This information and the corresponding information for the $i - j$ pair are used to calculate the pair-to-pair alignment score for each overhang. The best score, after subtracting the gap penalty appropriate for the overhang, is assigned to the $i - j$ residue pair. This procedure finds the best nongapped alignment, but it does not guarantee finding the best global alignment when it contains gaps. A rigorous algorithm for finding the true globally optimum alignment using these pair-by-pair score matrices is not known at the present time.

Selection of the test set of proteins

To test the new score matrices, each of a set of probe sequences of known structure were “threaded” through each structure in the domain database. The test consisted of scoring how many of the target proteins (domains that are structurally homologous to the probe protein) could be identified for each probe sequence. The probe sequences selected were the first members of each family in the October 1996 release of the SCOP database (Murzin et al., 1995), which were between 60 and 350 residues in length and which had at least one structural homologue of low sequence homology in the domain database.

We used two different criteria for deciding whether a probe and a domain were to be considered structurally homologous. For the purpose of selecting the probe sequences, a domain was considered to be a structural homologue if the number of matched residues after the structure–structure alignment by SHEBA was greater than 50% of the size of the *larger* protein. This tended to select only single domain proteins as the probe sequence. For the purpose of determining true and false positives after threading, the criterion used was that the number of matched residues by SHEBA alignment was greater than 60% of the *smaller* protein. This second criterion was used to recognize partial, as well as full, structural matches.

The sequence homology was considered to be low if the number of identities was <30% of the structurally aligned residues. The names of the 107 probe sequences selected by this procedure are given in Table 10.

BLAST and PSI-BLAST runs

The BLAST runs were made for each test protein sequence against the 3,539 structures in the domain database. The PSI-BLAST runs were made for each test sequences against the Swiss-Prot protein sequence database. A BLAST run was then made for each of the hits obtained against all the PDB sequences to identify the hit sequences for which the structure is known. The Blosum62 matrix and default settings were used for both procedures. The number of false positive BLAST hits, using the same SHEBA criteria for the structural similarity, was 3% for the BLAST and 0.8% for the PSI-BLAST runs. The low false positive rate for the PSI-BLAST run is probably due to the fact that the second stage BLAST run in this procedure was run against all PDB structures rather than against the domain database. Running against all PDB increases the number of correct but essentially duplicate hits, which in turn reduces the false positive rate. The numbers plotted in the inset of Figure 6 include only those for which there was at least one true positive, as is the case for all other procedures.

Acknowledgments

We thank Professor W. Pearson for making the FASTA program freely available and Dr. Shrish Tiwari for running the BLAST and PSI-BLAST tests. We also thank Mr. Joe J. Cammisa for his assistance with the computer system support and other members of the Laboratory for numerous discussions and criticisms. Jongsun Jung thanks Professors Frederick W. Carson, Lou Thompson Hughes, and Nina M. Roscher of the American University for their encouragement.

References

Altschul SF. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol 219*:555–565.

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol 215*:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BALST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*:3389–3402.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–169.

Brenner SE, Chothia C, Hubbard TJ. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA 95*:6073–6078.

Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Steven A B. 1997. An analysis of simultaneous variation in protein structures. *Protein Eng 10*:307–316.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: MO Dayhoff, ed. *Atlas of protein sequence and structure*. Vol. 5, Suppl. 3. Washington, DC: National Biomedical Research Foundation. pp 345–358.

Di Francesco V, Geetha V, Garnier J, Munson PJ. 1997. Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins Suppl 1*:123–128.

Eisenberg D, Luthy R, Bowie JU. 1997. VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol 277*:396–404.

Fischer D, Rice D, Bowie JU, Eisenberg D. 1996. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J 10*:126–136.

Flockner H, Domingues FS, Sippl MJ. 1997. Protein folds from pair interactions: A blind test in fold recognition. *Proteins Suppl*:129–133.

Göbel U, Sander C, Schneider R, Valencia A. 1994. Correlated mutations and residue contacts in proteins. *Proteins 18*:309–317.

Gonnet GH, Cohen MA, Benner SA. 1992. Exhaustive matching of the entire protein sequence database. *Science 256*:1443–1445.

Gonnet GH, Cohen MA, Benner SA. 1994. Analysis of amino-acid substitution during divergent evolution—The 400 by 400 dipeptide substitution matrix. *Biochem Biophys Res Comm 199*:489–496.

Henikoff S, Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915–10919.

Henikoff S, Henikoff J. 1993. Performance evaluation of amino acid substitution matrices. *Proteins Struct Funct Genet 17*:49–61.

Holm L, Sander C. 1994. Parser for protein folding units. *Proteins 19*:256–268.

Jaroszewski L, Rychlewski L, Zhang B, Godzik A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci 7*:1431–1440.

Jones D, Taylor W, Thornton J. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS 8*:275–282.

Jones DT. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol 287*:797–815.

Jung J, Lee B. 2000. Comparison of protein structures using an initial alignment based on environmental profile. *Protein Eng 13*. Forthcoming.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Karlin S, Altschul SF. 1991. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. 1997. Predicting protein structure using hidden Markov models. *Proteins Suppl 1*:134–139.

Koehl P, Levitt M. 1999. A brighter future for protein structure prediction. *Nat Struct Biol 6*:108–111.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Levitt M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins Suppl*:92–104.

Marchler-Bauer A, Bryant SH. 1997. Measures of threading specificity and accuracy. *Proteins Suppl*:74–82.

Marchler-Bauer A, Levitt M, Bryant SH. 1997. A retrospective analysis of CASP2 threading predictions. *Proteins Suppl 1*:83–91.

Miller RT, Jones DT, Thornton JM. 1996. Protein fold recognition by sequence threading: tools and assessment techniques. *FASEB J 10*:171–178.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequence and structures. *J Mol Biol 247*:536–540.

Nakayama S, Shigezumi S, Yoshida M. 1988. Method for clustering proteins by use of all possible pairs of amino acids as structural descriptors. *J Chem Inf Comput Sci 28*:72–78.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

Neher E. 1994. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA 91*:98–102.

Olmea O, Valencia A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des 2*:S25–32.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino-acid substitution tables—Tertiary templates and prediction of protein folds. *Protein Sci 1*:216–226.

Pearson WR. 1998. Empirical statistical estimates for sequence similarity searches. *J Mol Biol 276*:71–84.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444–2448.

Rice DW, Eisenberg D. 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol 267*:1026–1038.

Rice DW, Fischer D, Weiss R, Eisenberg D. 1997. Fold assignments for amino acid sequences of the CASP2 experiment. *Proteins Suppl 1*:113–122.

Russell RB, Saqi MA, Bates PA, Sayle RA, Sternberg MJ. 1998. Recognition of analogous and homologous protein folds—Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng 11*:1–9.

Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng 7*:349–358.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195–197.

Taylor WR, Hatrick K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng 7*:341–348.

van Heel M. 1991. A new family of powerful multivariate statistical sequence analysis. *J Mol Biol 220*:877–887.